Assignment 5: Data Visualization

Yared S. Asfaw

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Visualization

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_A05_DataVisualization.Rmd") prior to submission.

The completed exercise is due on Monday, February 14 at 7:00 pm.

Set up your session

1. Set up your session. Verify your working directory and load the tidyverse and cowplot packages. Upload the NTL-LTER processed data files for nutrients and chemistry/physics for Peter and Paul Lakes (use the tidy [`NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv`] version) and the processed data file for the Niwot Ridge litter dataset (use the [`NEON_NIWO_Litter_mass_trap_Processed.csv`] version).

```
# Locating the working directory
getwd()
```

```
## [1] "E:/EDA/Environmental_Data_Analytics_2022"
```

```
# Loading packages
library(tidyverse)
#install.packages("cowplot")
library(cowplot)
library(ggplot2)
```

```
# Importing the dataset
ntl_nutrients <- read.csv("./Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv",
                          stringsAsFactors = TRUE)
niwo_litter <- read.csv("./Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv",
                        stringsAsFactors = TRUE)
```

2. Make sure R is reading dates as date format; if not change the format to date.

```
#1 Checking and formatting the date

class(ntl_nutrients$sampledate) # The sample date is "factor" and it is in ymd format
```

```
## [1] "factor"
```

```
ntl_nutrients$sampledate <- as.Date(ntl_nutrients$sampledate, format= "%Y-%m-%d")
```

```
class(ntl_nutrients$sampledate)
```

```
## [1] "Date"
```

```
class(niwo_litter$collectDate) # The sample collect date is "factor" and it is in ymd format
```

```
## [1] "factor"
```

```
niwo_litter$collectDate <- as.Date(niwo_litter$collectDate, format = "%Y-%m-%d")
```

```
class(niwo_litter$collectDate)
```

```
## [1] "Date"
```

Define your theme

3. Build a theme and set it as your default theme.

```
# Defining a theme named "mytheme"
mytheme <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black"), legend.position ="bottom")
```

```
# Setting the theme created as default
theme_set(mytheme)
```

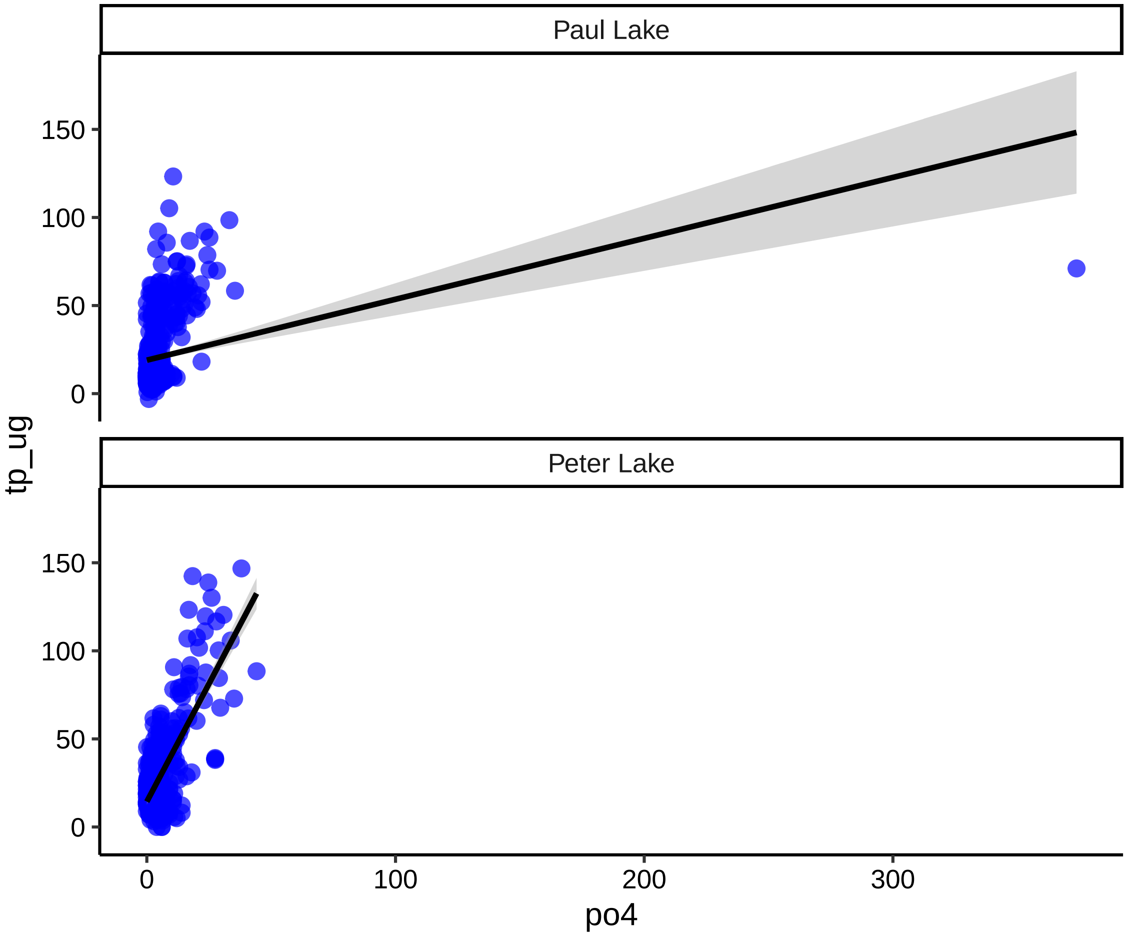
Create graphs

For numbers 4-7, create ggplot graphs and adjust aesthetics to follow best practices for data visualization. Ensure your theme, color palettes, axes, and additional aesthetics are edited accordingly.

4. [NTL-LTER] Plot total phosphorus (`tp_ug`) by phosphate (`po4`), with separate aesthetics for Peter and Paul lakes. Add a line of best fit and color it black. Adjust your axes to hide extreme values (hint: change the limits using `xlim()` and `ylim()`).

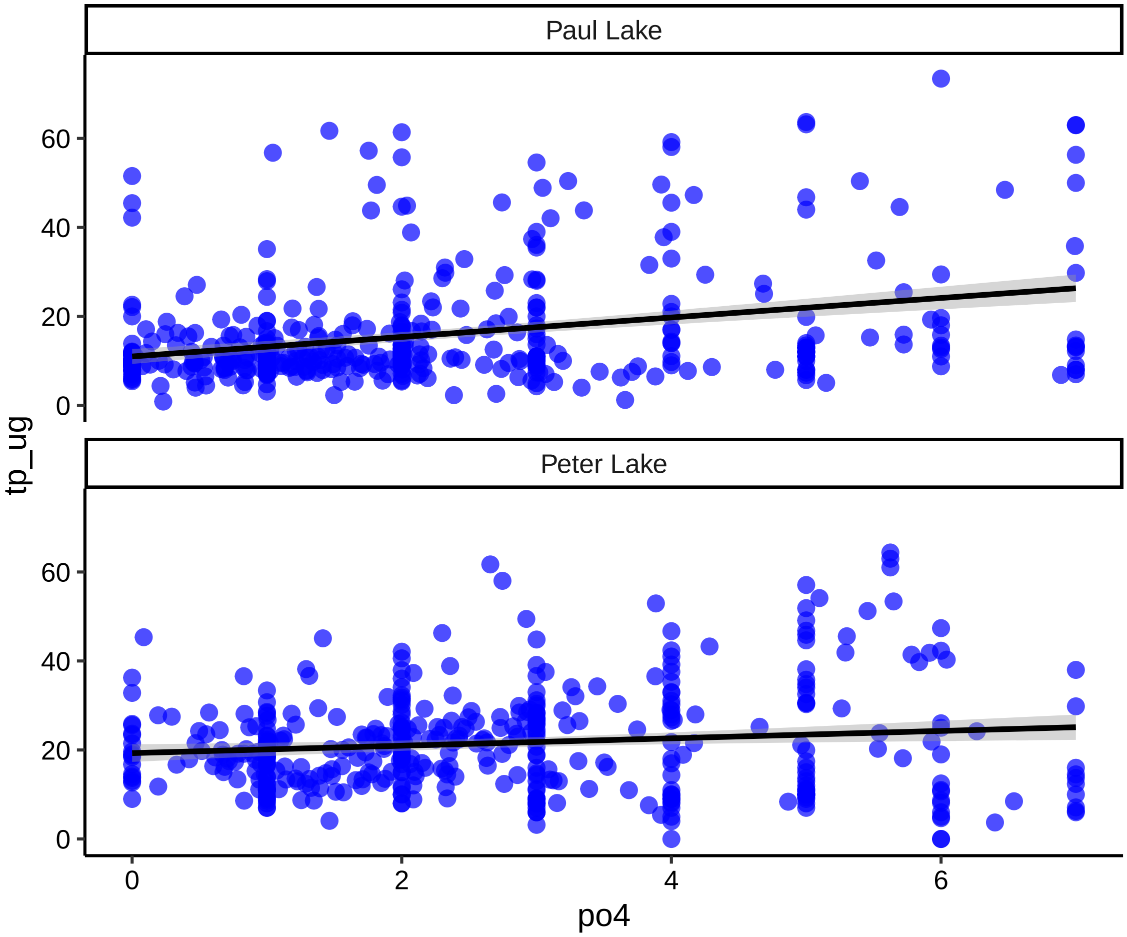
```
# Total phosphorus (tp_ug) by phosphate ('po4') plot
ggplot(ntl_nutrients, aes(x=po4, y=tp_ug)) +
  facet_wrap(vars(lakename), nrow = 2) +
  geom_point(alpha = 0.7, size = 2.5, color="blue")+
  geom_smooth(method="lm", color="black")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



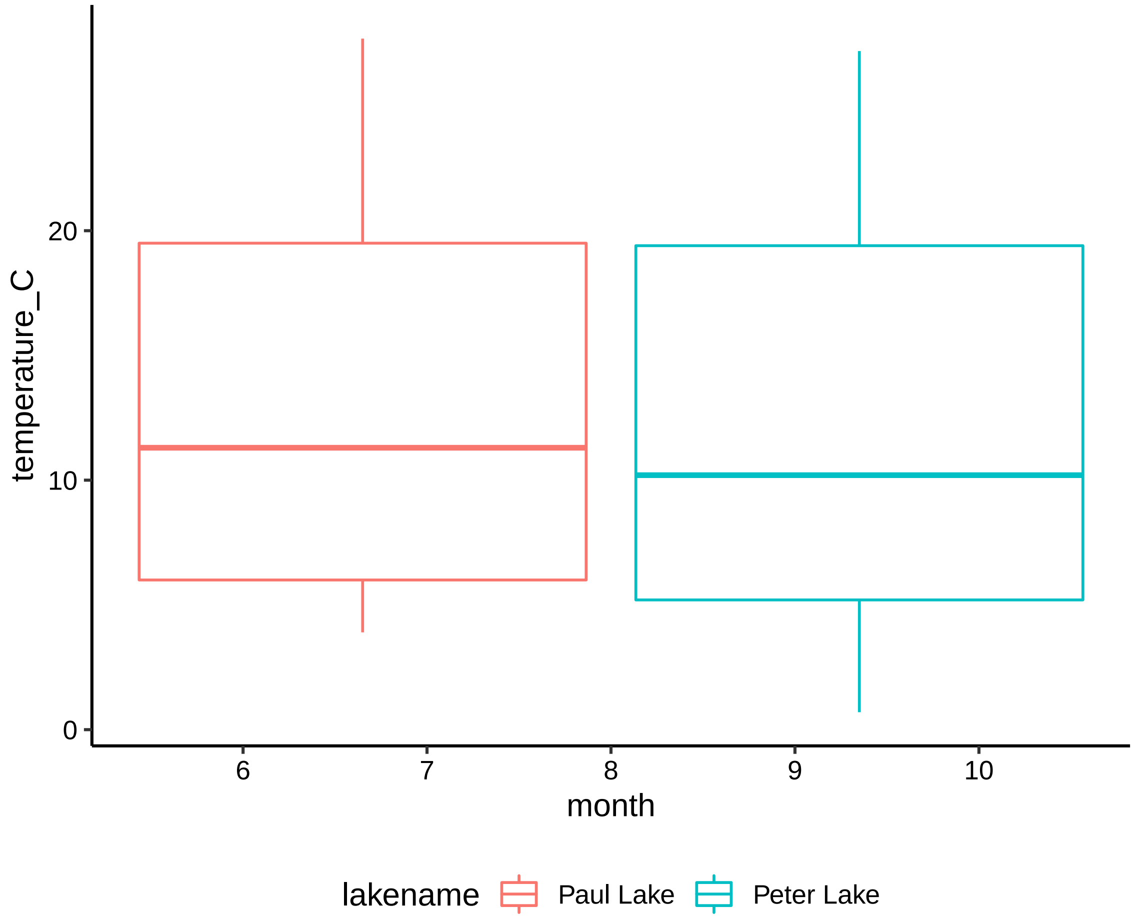
```
# Adjusting the axes to hide extreme values Total phosphorus (tp_ug) by phosphate ('po4') plot
ggplot(ntl_nutrients, aes(x=po4, y=tp_ug)) +
  facet_wrap(vars(lakename), nrow = 2) +
  geom_point(alpha = 0.7, size = 2.5, color="blue")+
  geom_smooth(method="lm", color="black") +
  mytheme +
  xlim(0,7) +
  ylim(0,75)
```

```
## 'geom_smooth()' using formula 'y ~ x'
```

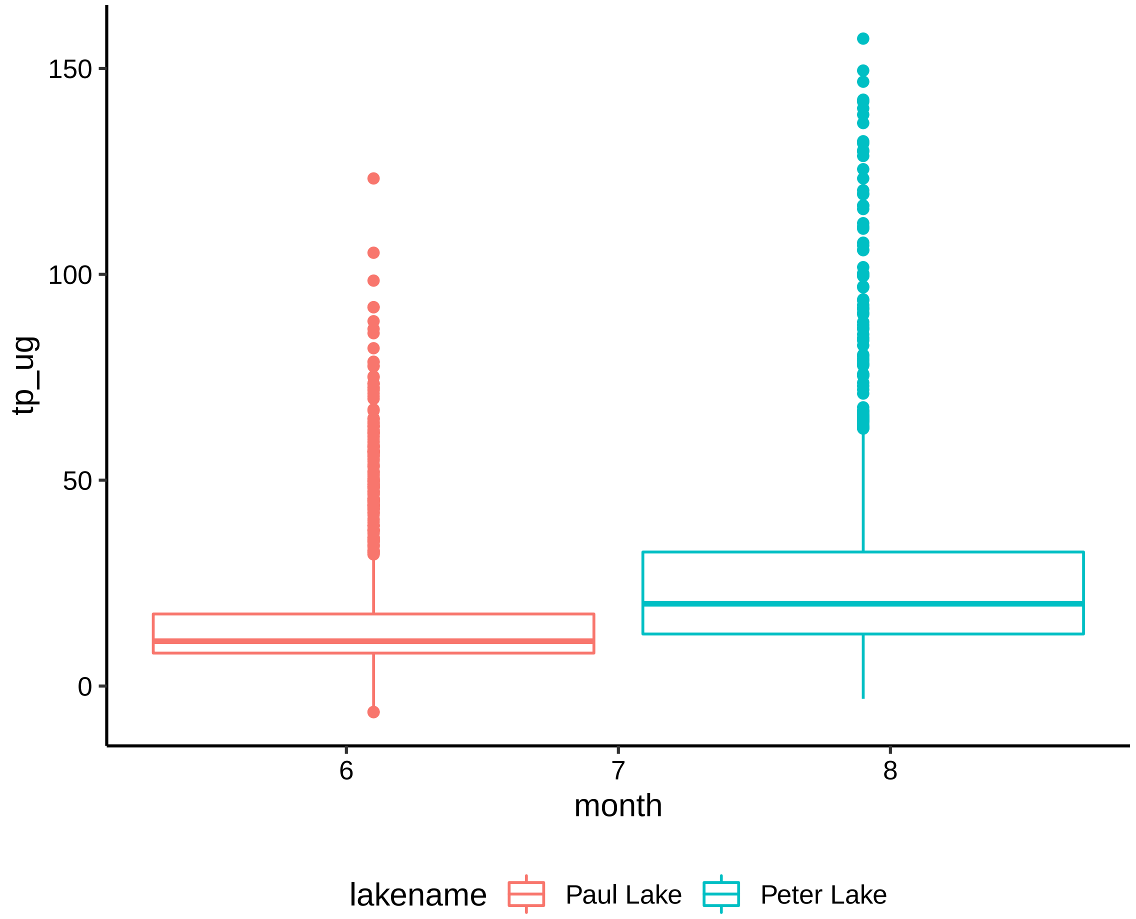


5. [NTL-LTER] Make three separate boxplots of (a) temperature, (b) TP, and (c) TN, with month as the x axis and lake as a color aesthetic. Then, create a cowplot that combines the three graphs. Make sure that only one legend is present and that graph axes are aligned.

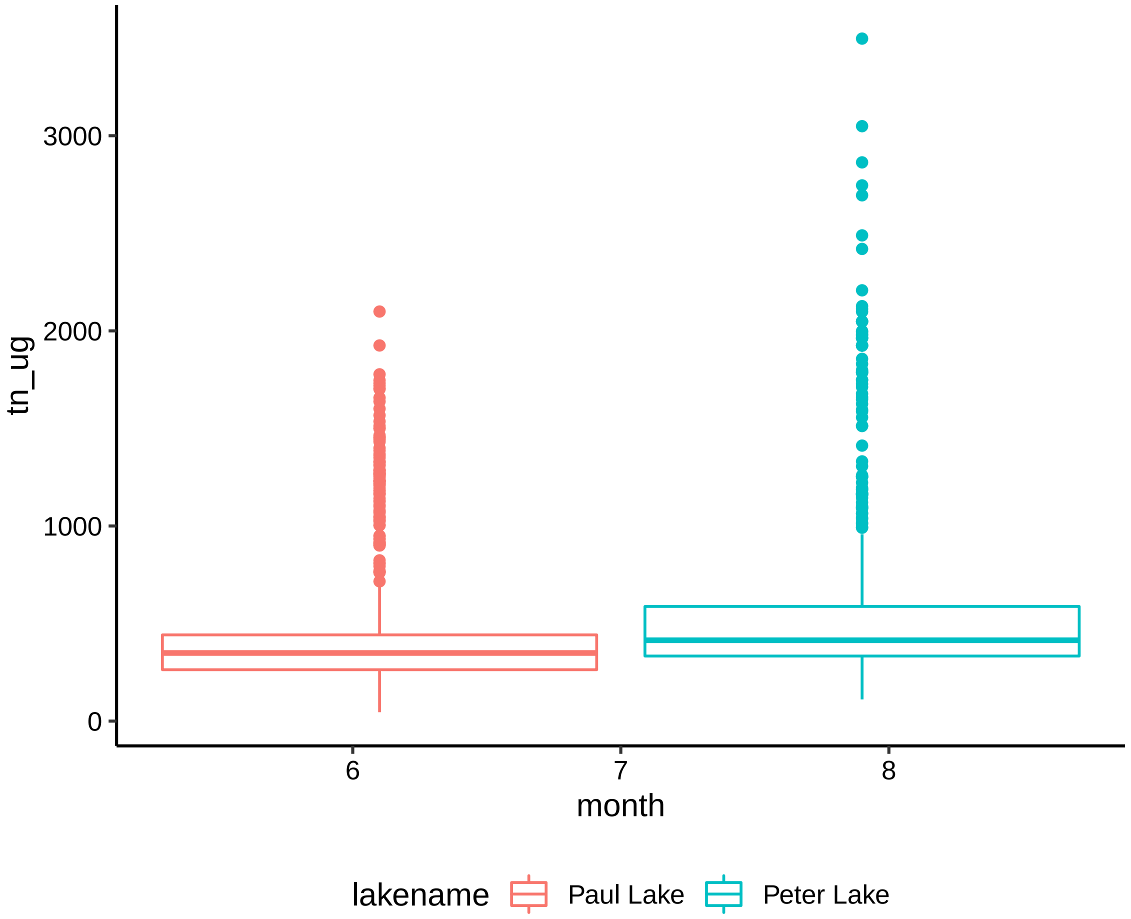
```
# Plots for temperature, TP and TN
temp.plot <- ggplot(ntl_nutrients, aes(x=month, y= temperature_C)) +
  geom_boxplot(aes(color=lakename))
print(temp.plot)
```



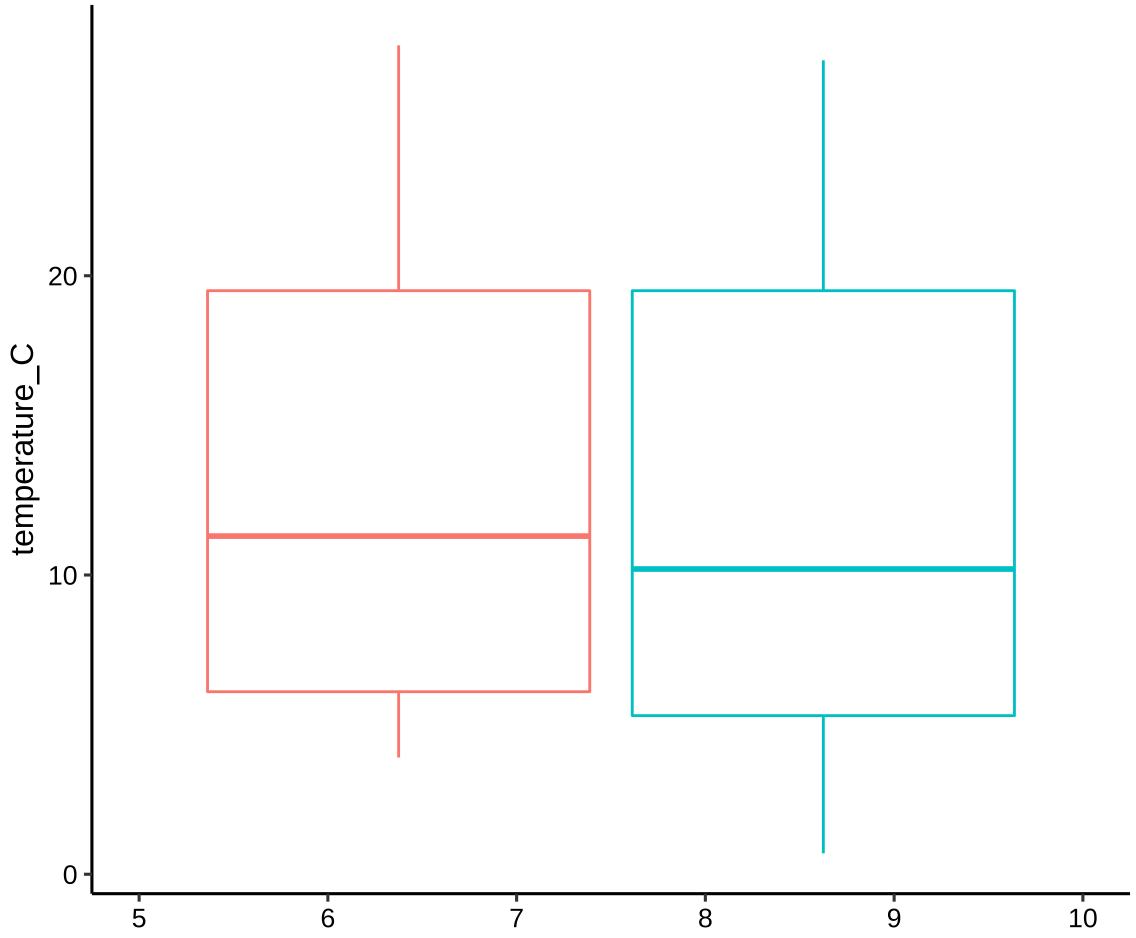
```
tp.plot <- ggplot(ntl_nutrients, aes(x=month, y= tp_ug)) +
  geom_boxplot(aes(color=lakename))
  print(tp.plot)
```



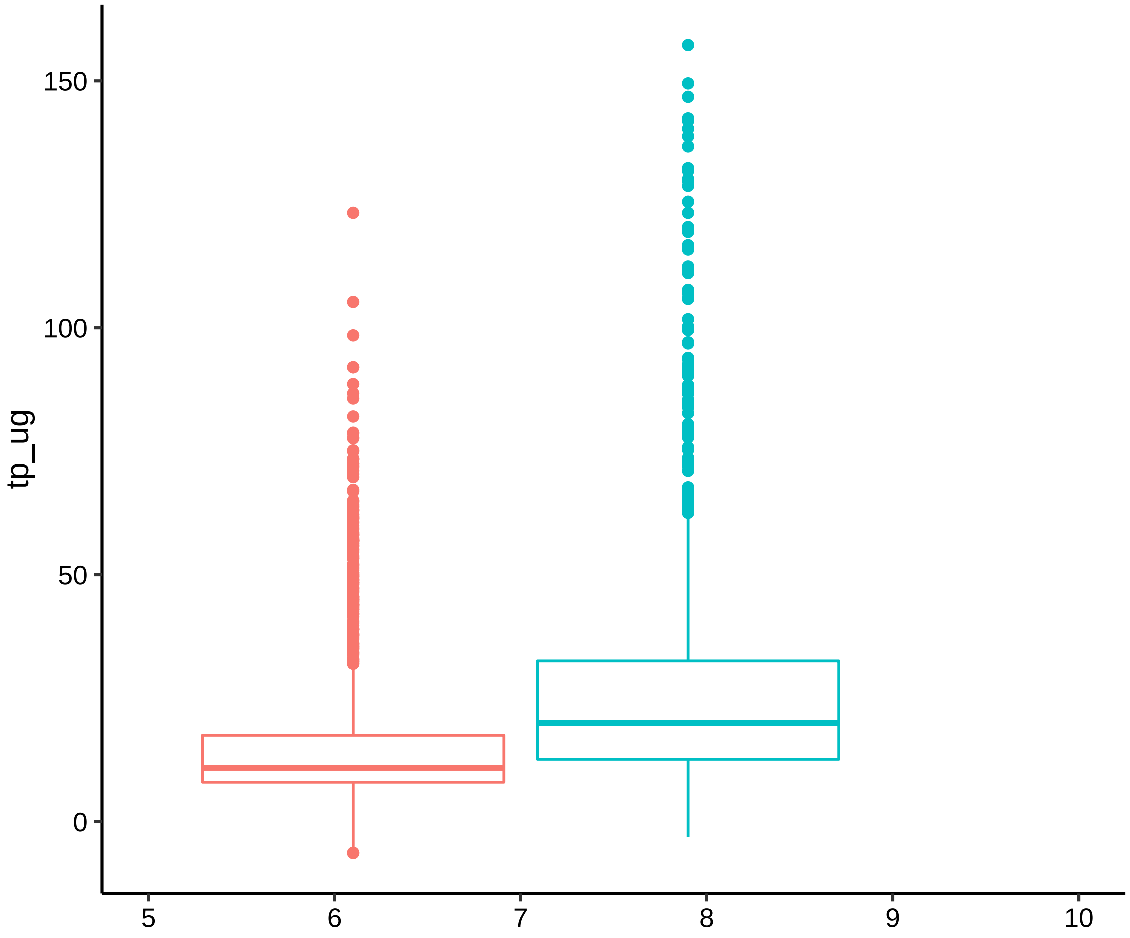
```
tn.plot <- ggplot(ntl_nutrients, aes(x=month, y= tn_ug)) +
  geom_boxplot(aes(color=lakename))
print(tn.plot)
```



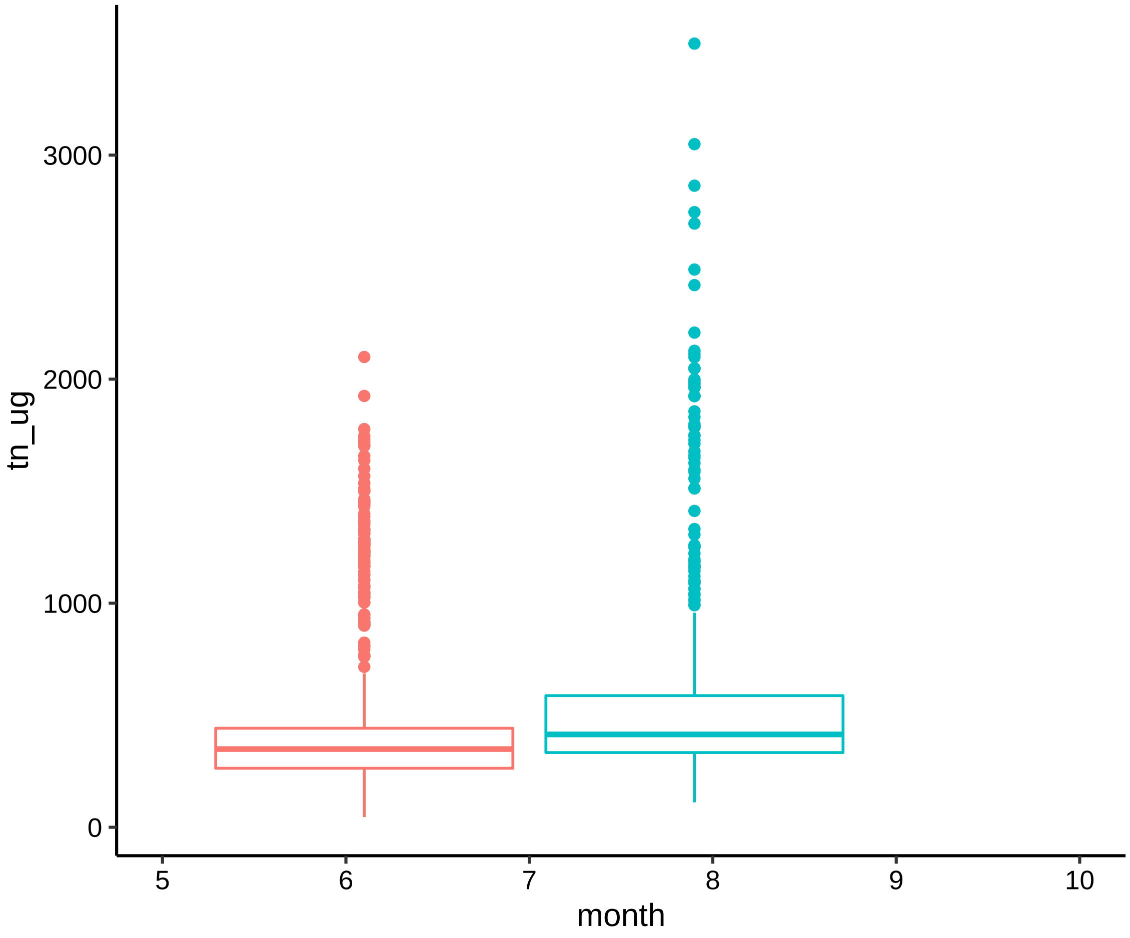
```
# Create all the three plots without legend and x-axis lable (except the x-axis
# lable of tn.plot or the last one)
# Extend the limit of the x axis to align the plots vertically
temp.plot <- ggplot(ntl_nutrients, aes(x=month, y= temperature_C)) +
  geom_boxplot(aes(color=lakename)) +
    xlim(5,10) +
    xlab(NULL) +
  theme(legend.position = "none")
print(temp.plot)
```



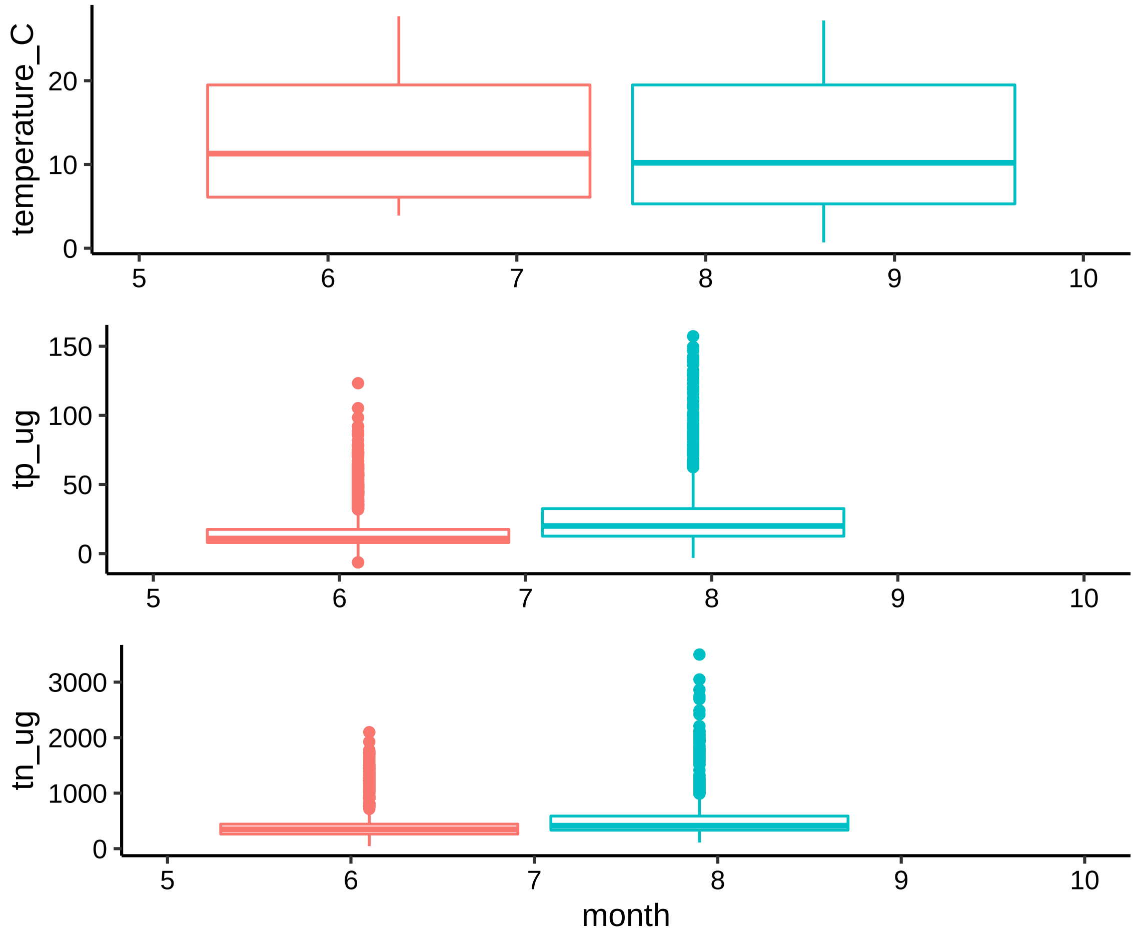
```
tp.plot <- ggplot(ntl_nutrients, aes(x=month, y= tp_ug)) +
  geom_boxplot(aes(color=lakename)) +
  xlim(5,10) +
  xlab(NULL) +
  theme(legend.position = "none")
print(tp.plot)
```



```
tn.plot <- ggplot(ntl_nutrients, aes(x=month, y= tn_ug)) +
    geom_boxplot(aes(color=lakename)) +
    xlim(5,10) +
    theme(legend.position = "none")
print(tn.plot)
```

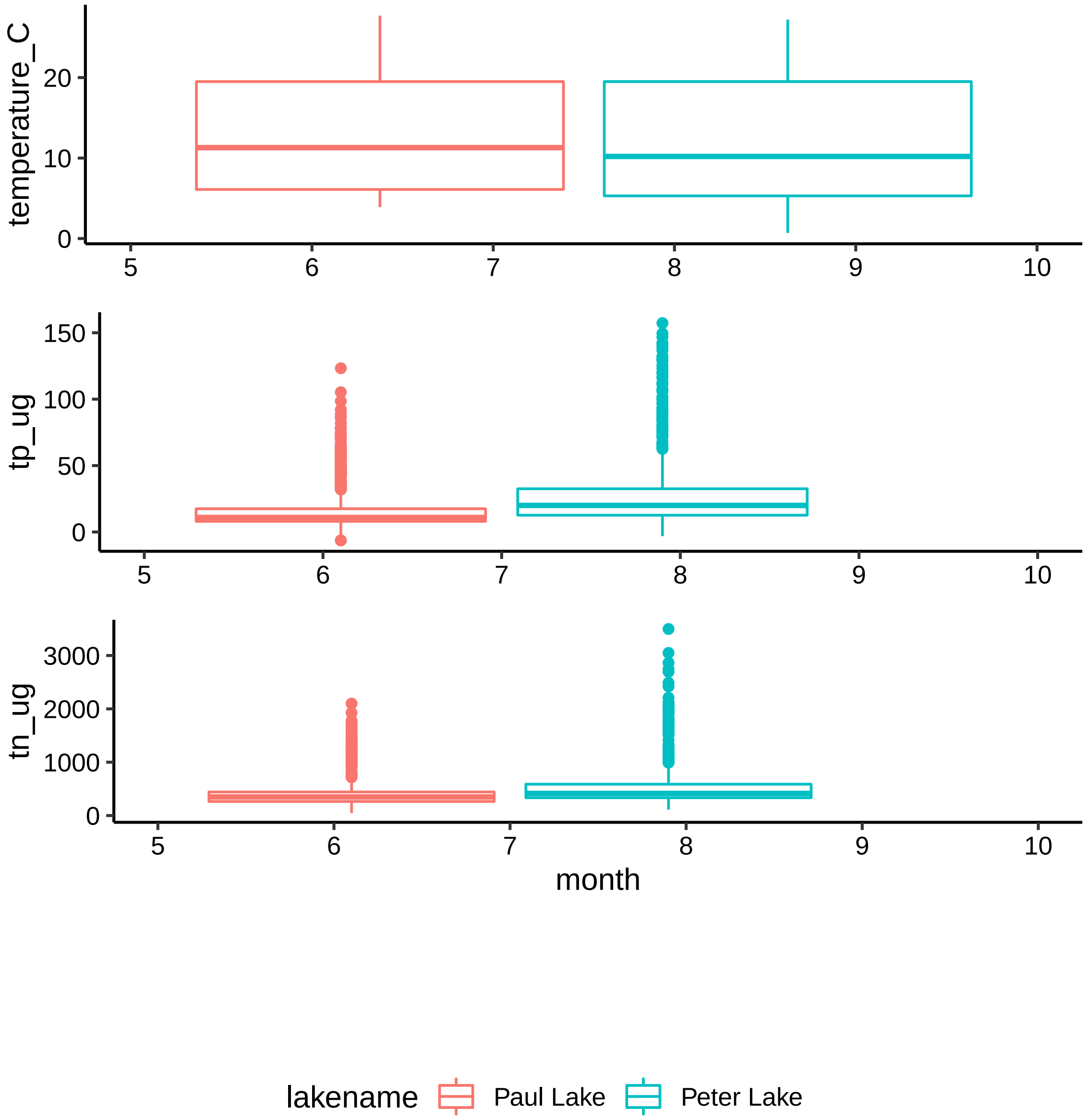


```
# Combine the three plots in one column
combined_plots <- plot_grid(temp.plot, tp.plot, tn.plot, ncol = 1)
print(combined_plots)
```



```
# extract legend from one of the plots
legend <- get_legend(temp.plot + theme(legend.position = "bottom"))
```

```
# Combine the combined_plots and legend
plot_grid(combined_plots, axis = "b", legend, ncol = 1, align = 'v', rel_heights = c(1.25,0.5,0.5))
```



Question: What do you observe about the variables of interest over seasons and between lakes?

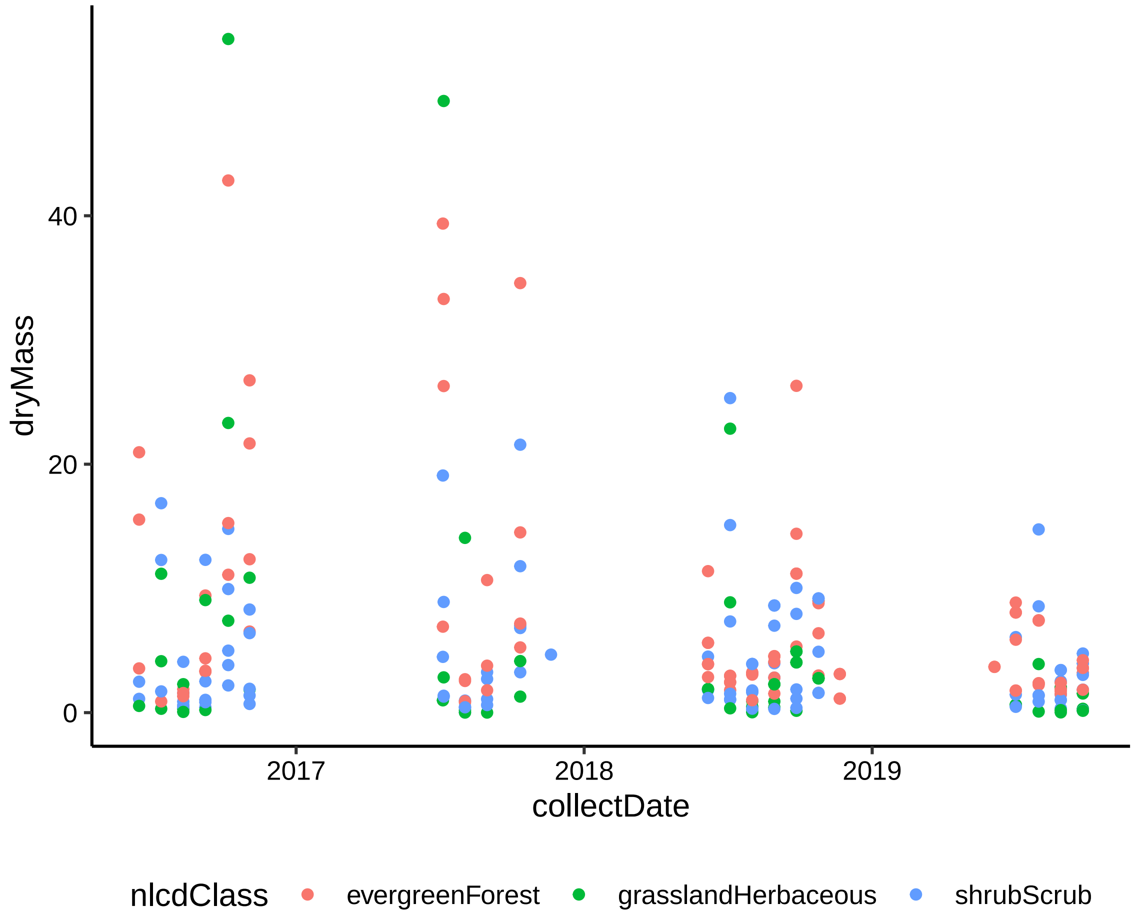
> Answer: In both lakes most of the temperature observations fall above the median values with Peter lake having a higher number of observations above the median value; and the variablity of the temprature observations is nearly the same for both lakes with some level of higher variblity in Peter lake.
> Peter Lake: There is high variablity of tp_ug and tn_ug concentration in Peter lake with the tp_ug having a higher number of observations above the median value, otherwise in both cases higher number of observations lie above the median value. Extreme values or outliers are also nearly the same in both tp_ug and tn_ug cases. Paul Lake: There is a higher variablity of tp_ug than tn_ug concenrations. In case of tp_ug higher number of observations fall above the

median value whereas in tn_ug the higher number of observations fall below the median value. In both tp_ug and tn_ug cases, there are extreme values (outliers) with tp_ug having the higher numebr of outliers.

The total phosphurus (tp_nu) and total nitrogen (tn_ug) concentrations are higher in Peter lake than Paul lake whereas Paul lake has a slightly higher temperature than Peter lake. In both lakes, there are significant number of extreme nutrient concentrations (outliers) that are higher than third quartile value but again this is much higher for Peter lake.

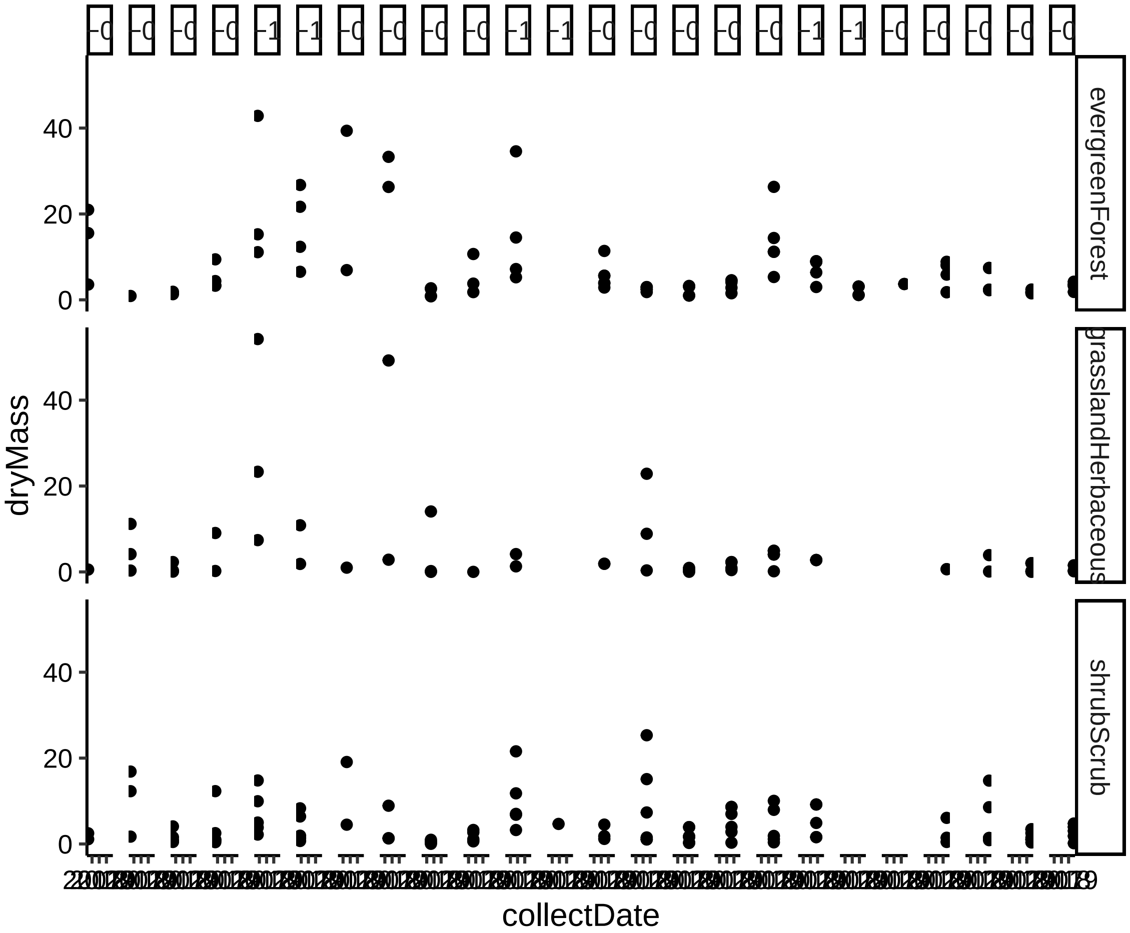
6. [Niwot Ridge] Plot a subset of the litter dataset by displaying only the "Needles" functional group. Plot the dry mass of needle litter by date and separate by NLCD class with a color aesthetic. (no need to adjust the name of each land use)

```r
# Plotting the dry mass of needle litter
ggplot(subset(niwo_litter,functionalGroup == "Needles"), aes(x=collectDate, y=dryMass))+
  geom_point(aes(color=nlcdClass))
```



7. [Niwot Ridge] Now, plot the same plot but with NLCD classes separated into three facets rather than separated by color.

```
# Plotting the dry mass of needle litter with three facets of nlcd classes
ggplot(subset(niwo_litter,functionalGroup == "Needles"), aes(x=collectDate, y=dryMass))+
  geom_point()+
  facet_grid(nlcdClass ~ collectDate)
```



Question: Which of these plots (6 vs. 7) do you think is more effective, and why?

> Answer: The plot under 7 is easier to examine the individual classes as it is more effective in easily conveying the information of each class to the user. The three classes with the corresponding years help to easily understand the drymass situation of those land use systems.However, in the plot under 6, it is easier to see how the classes are clustered overall. So, depending on the size of the dataset and the information we would like to draw, either options can be effective. In this case, considering the dataset as large, we are more likely interested in the overall clustering instead of the individual point clouds. Thus, the plot under 6 is more effective in demonstrating that as it differentiates the classes in color and show their cluster.